RELIABLE *A POSTERIORI* SIGNAL-TO-NOISE RATIO FEATURES SELECTION

Cyril Plapous [1], *Claude Marro* [1], *Pascal Scalart* [2]

[1] France Télécom - TECH/SSTP, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
[2] University of Rennes - IRISA / ENSSAT, 6 Rue de Kerampont, B.P. 80518, 22305 Lannion, France
E-mail: cyril.plapous,claude.marro@francetelecom.com; pascal.scalart@enssat.fr

ABSTRACT

This paper addresses the problem of single microphone speech enhancement in noisy environments. State of the art short-time noise reduction techniques are most often expressed as a spectral gain depending on the Signal-to-Noise Ratio (SNR). The well-known decision-directed (DD) approach drastically limits the level of musical noise but the estimated *a priori* SNR is biased since it depends on the speech spectrum estimated in the previous frame. The consequence of this bias is an annoying reverberation effect. We propose a new method, called Reliable Features Selection Noise Reduction (RFSNR) technique, that is able to classify the *a posteriori* SNR estimates into two categories: the reliable features leading to speech components and the unreliable ones corresponding to musical noise only. Then it is possible to directly enhance speech using *a posteriori* SNR leading to an unbiased estimator.

1. INTRODUCTION

The problem of enhancing speech degraded by additive noise, when only the noisy speech is available, has been widely studied in the past and is still an active field of research. Noise reduction is useful in many applications such as voice communication and automatic speech recognition.

Scalart and Vieira Filho presented in [1] an unified view of the main single microphone noise reduction techniques where the process relies on the estimation of a short-time spectral gain which is a function of the *a priori* Signal-to-Noise Ratio (SNR) and/or the *a posteriori* SNR. They also emphasize the interest of estimating the *a priori* SNR with the decision-directed (DD) approach proposed by Ephraim and Malah in [2]. Cappé analyzed the behavior of this estimator in [3] and demonstrated that the *a priori* SNR follows the shape of the *a posteriori* SNR with a one frame delay. Consequently, since the gain depends on the *a priori* SNR, it does not match anymore the current frame and thus it degrades the performance of the noise reduction system.

We propose a method, called Reliable Features Selection Noise Reduction (RFSNR) technique, that uses the *a priori* SNR estimated with the DD approach and the *a posteriori* SNR in order to classify this latter into reliable or unreliable features. This approach allows an efficient separation of speech components from musical noise ones. Indeed, the enhanced speech is obtained using unbiased SNR estimator and is free of musical noise.

2. CLASSICAL DECISION-DIRECTED APPROACH

2.1. Noise reduction parameters

In the classical additive noise model, the noisy speech is given by $x(t) = s(t) + n(t)$ where $s(t)$ and $n(t)$ denote the speech and the noise signal, respectively. Let $S(p,k)$, $N(p,k)$ and $X(p,k)$ designate the kth spectral component of short-time frame p of the speech $s(t)$, the noise $n(t)$ and the noisy speech $x(t)$, respectively.

The objective is to find an estimator $\hat{S}(p,k)$ which minimizes the expected value of a given distortion measure conditionally to a set of spectral noisy features. Since the statistical model is generally nonlinear, and since there does not exist any simple solution for the spectral estimation, we first derive an SNR estimate from the noisy features. An estimate of $S(p,k)$ is subsequently obtained by applying a spectral gain $G(p,k)$ to each short-time spectral component $X(p,k)$. This gain corresponds to different functions proposed in the literature (*e.g.* amplitude and power spectral subtraction, Wiener filtering, MMSE STSA, *etc.*) [4, 5, 1, 2]. The choice of the distortion measure determines the gain behavior, *i.e.* the well-known trade-off between noise reduction and speech distortion. However, the key parameter is the estimated SNR because it determines the efficiency of the speech enhancement for a given noise power spectrum density (PSD).

Most of the classical speech enhancement techniques require the evaluation of two parameters, the *a posteriori* SNR and the *a priori* SNR, respectively defined by

$$SNR_{post}(p,k) = \frac{|X(p,k)|^2}{\mathbb{E}[|N(p,k)|^2]} \tag{1}$$

and
$$SNR_{prio}(p,k) = \frac{\mathbb{E}[|S(p,k)|^2]}{\mathbb{E}[|N(p,k)|^2]}, \tag{2}$$

where $\mathbb{E}[.]$ is the expectation operator. In practical implementations, the PSDs of speech $|S(p,k)|^2$ and noise $|N(p,k)|^2$ are unknown as only the noisy speech is available, then both SNRs have to be estimated. The estimation of the noise PSD, $\hat{\gamma}_{nn}(p,k)$, is beyond our scope and can be easily computed during speech pauses using recursive averaging.

2.2. Decision-Directed approach

Generally, the two estimated SNRs are computed as follows

$$\hat{SNR}_{post}(p,k) = \frac{|X(p,k)|^2}{\hat{\gamma}_{nn}(p,k)} \tag{3}$$

and $\hat{SNR}_{prio}(p,k) = \beta \frac{|\hat{S}(p-1,k)|^2}{\hat{\gamma}_{nn}(p,k)}$
$$+(1-\beta)P[\hat{SNR}_{post}(p,k)-1] \tag{4}$$

where $P[.]$ denotes the half-wave rectification and $\hat{S}(p-1,k)$ is the estimated speech spectrum at previous frame. This *a priori*

SNR estimator corresponds to the so-called decision-directed approach [2, 3] whose behavior is controlled by the parameter β (typically $\beta = 0.98$). The approaches based on (3) and (4) to compute the spectral gain will be referred to the DD algorithm.

We can emphasize two effects of the DD algorithm which have been interpreted by Cappé in [3]:

- When the *a posteriori* SNR is much larger than 0dB, $\hat{SNR}_{prio}(p,k)$ corresponds to a one frame delayed version of $\hat{SNR}_{post}(p,k) - 1$.
- When the *a posteriori* SNR is lower or close to 0dB, $\hat{SNR}_{prio}(p,k)$ corresponds to a highly smoothed and delayed version of $\hat{SNR}_{post}(p,k) - 1$. The direct consequence for the enhanced speech is the reduction of the musical noise effect due to a lower variance.

The delay inherent to the DD algorithm is a drawback especially during the speech non-stationarities like speech onset and offset. Furthermore, this delay introduces a bias in the gain estimation which limits the noise reduction performance and generates an annoying reverberation effect.

3. SNR ANALYSIS TOOL

In order to evaluate the behavior of speech enhancement techniques, we propose to use an approach described by Renevey and Drygajlo [6]. The basic principle is to consider the *a priori* SNR versus the *a posteriori* SNR in order to analyze the behavior of the features defined by the 2-tuple (SNR_{post}, SNR_{prio}).

In the additive model, the amplitude of the noisy signal can be expressed as $|X(p,k)| =$

$$\sqrt{|S(p,k)|^2 + |N(p,k)|^2 + 2|S(p,k)||N(p,k)|\cos\alpha(p,k)} \quad (5)$$

where $\alpha(p,k)$ is the phase difference between $S(p,k)$ and $N(p,k)$. The local *a posteriori* and *a priori* SNRs, assuming the knowledge of the clean speech and the noise, can be defined by

$$SNR^{local}_{post}(p,k) = \frac{|X(p,k)|^2}{|N(p,k)|^2} \quad (6)$$

and

$$SNR^{local}_{prio}(p,k) = \frac{|S(p,k)|^2}{|N(p,k)|^2}. \quad (7)$$

By replacing $|X(p,k)|$ in (6) by its expression (5) and using (7), it comes $SNR^{local}_{post}(p,k) =$

$$SNR^{local}_{prio}(p,k) + 1 + 2\sqrt{SNR^{local}_{prio}(p,k)}\cos\alpha(p,k). \quad (8)$$

This relation depends on $\alpha(p,k)$ which is an uncontrolled parameter in speech enhancement techniques. For example, in the derivation of the classical Wiener filter [1], the $SNR_{post}(p,k)$ is assumed to be equal to $SNR_{prio}(p,k) + 1$ which corresponds to a constant phase difference $\alpha(p,k) = \frac{\pi}{2}$ (*i.e.* noise and clean speech are supposed to be added in quadrature).

In the following, the discussion will be illustrated using a French sentence corrupted by car noise at 12dB global SNR but it can be generalized to other noise and SNR conditions. The spectrogram of this noisy sentence is shown in Fig. 4.(a). The relation expressed by (8) is illustrated in Fig. 1. The dark gray features represent the *a priori* SNR versus the *a posteriori* SNR in the ideal case where the clean speech and the noise amplitudes are known.

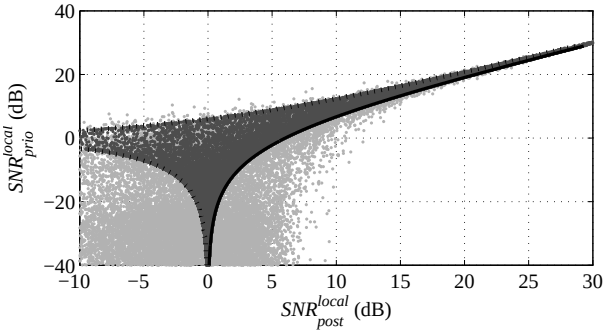

Figure 1: SNR^{local}_{prio} versus SNR^{local}_{post}. Dark gray features: clean speech and noise amplitudes are known in (6) and (7). Light gray features: clean speech amplitude is known but estimated noise PSD is used in (6) and (7).

The features lie between two curves, the solid one (resp. dashed) corresponds to the limit case where $\alpha(p,k) = 0$ (π), where noise and clean speech spectral components are added in phase (phase opposition). These two limits define an area where the feature repartition depends on the true phase difference $\alpha(p,k)$. Notice that since only the amplitudes of the signals are used to compute the SNRs involved in the spectral gain computation, estimation errors inherent to the speech enhancement method cannot be avoided even knowing the signals.

The light gray features in Fig. 1 represent the case where an estimation of the noise PSD is used in (6) and (7) instead of the local noise but still assuming the knowledge of the clean speech amplitude. Notice that in that case, the SNR^{local}_{post} corresponds to $\hat{SNR}_{post}$ of (3). The errors which occur in the noise PSD estimation lead to an important dispersion of the features outside of the limit area for low SNR values and decrease the quality of the enhanced speech.

4. RELIABLE *A POSTERIORI* SNR FEATURES

4.1. Comparison between *a posteriori* and *a priori* SNRs

It is interesting to underline the behavior of the *a posteriori* and *a priori* SNR estimators. It is well known that using only the *a posteriori* SNR to enhance the noisy speech results in a very high level of musical noise, leading to a very poor global quality signal. However, this is the technique leading to the lower degradation level for the speech components themselves. The *a priori* SNR, estimated in the DD approach, is widely used instead of the *a posteriori* SNR because the musical noise is reduced to an acceptable level. However, this estimated SNR is biased leading to underestimation or overestimation of SNR components and then reducing performance during speech activity. From a subjective point of view, this bias which is related to the delay effect described in section 2 is perceived as a reverberation effect.

In order to measure the performance of SNR estimators, it is useful to compare the estimated SNR values to the true ones as shown in Fig. 2 where the estimated SNR is displayed versus the true SNR (equation (6) for Fig. 2.(a) and (7) for Fig. 2.(b)). The SNRs are plotted for 50 frames of speech activity to focus the analysis on the behavior of the SNR estimators for speech components. Figure 2.(a) illustrates the case where the *a posteriori* SNR is estimated using equation (3) and Fig 2.(b) the case where the *a priori* SNR is estimated using the DD approach given by equation (4). In

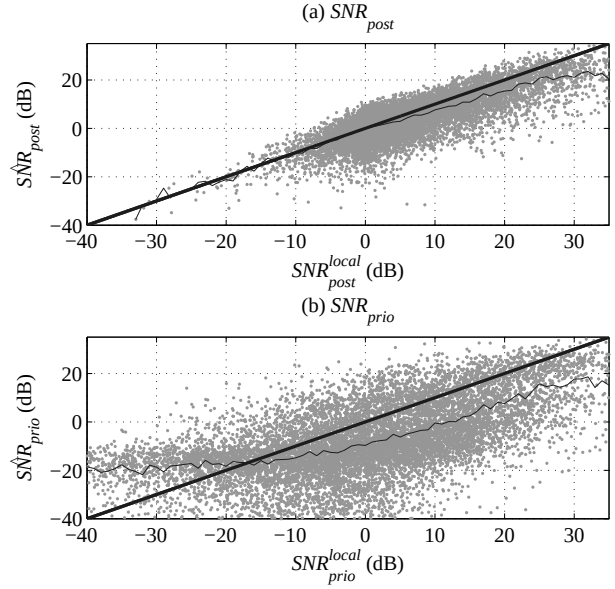

Figure 2: Estimated SNR versus true SNR (*i.e.* local SNR) in case of (a) *a posteriori* SNR and (b) *a priori* SNR. The bold line represents a perfect estimator and the thin line represents the mean of the estimated SNR versus the true SNR.

these two cases, the bold line corresponds to a perfect SNR estimator that can be used as a reference to evaluate the performance of the real estimators. It is obvious that the features corresponding to the *a posteriori* SNR estimator are closer to the reference bold line and less dispersed than the *a priori* SNR estimator ones.

The dispersion observed for the two cases (a) and (b) of Fig 2 can be characterized by the covariance which can be computed as $cov(S\hat{N}R; SNR) =$

$$\mathbb{E}\left[(S\hat{N}R - \mathbb{E}[S\hat{N}R])(SNR - \mathbb{E}[SNR])\right] \qquad (9)$$

where $S\hat{N}R$ and SNR denotes the estimated and true SNRs, respectively. For the typical cases depicted in Fig. 2, we obtain $cov\left(S\hat{N}R_{prio}; SNR_{prio}\right) \approx 2cov\left(S\hat{N}R_{post}; SNR_{post}\right)$, which corresponds to a greater dispersion for the *a priori* SNR.

In Fig. 2.(a) and (b), the thin line represents the mean of the estimated SNR knowing the true SNR and is obtained as follows

$$\mathbb{E}[S\hat{N}R|SNR] = \int s\hat{n}r \ \mathrm{p}(s\hat{n}r|SNR)\, ds\hat{n}r \qquad (10)$$

where p is the probability density function. The mean of the estimated SNR is closer to the perfect estimator for the *a posteriori* SNR estimator. It is slightly underestimated for high SNR whereas for the *a priori* SNR the underestimation is large for SNR greater than -17dB. However, since the dispersion is high for the *a priori* SNR features, even if the mean is largely underestimated, the case where SNR features are overestimated exists. Furthermore, the *a priori* SNR is overestimated for SNR smaller than -17dB. Finally, these results confirm that the *a posteriori* SNR estimator is more reliable than the *a priori* SNR estimator for speech components.

4.2. Reliable *a posteriori* SNR features selection

Since the *a posteriori* SNR estimator is better for speech components than the *a priori* SNR estimator of the DD approach, a judicious strategy would be to determine when it is possible to use it and when it will lead to musical noise. In order to select only the reliable *a posteriori* SNR components, we propose to separate the SNR features in the space defined by the 2-tuple $(S\hat{N}R_{post}, S\hat{N}R_{prio})$ using two thresholds. Given the threshold η for the *a priori* SNR, it is possible to compute the threshold δ for the *a posteriori* SNR using (8) which depends on the phase parameter $\alpha(p,k)$. As displayed in Fig. 3 these SNR features will be then separated into four quadrants. We propose to

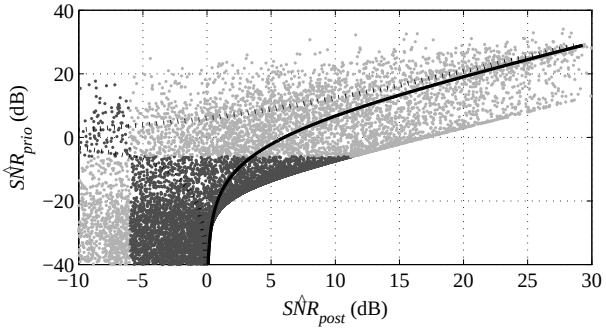

Figure 3: Separation of the features defined by $S\hat{N}R_{prio}$ of DD approach versus $S\hat{N}R_{post}$. The RFSNR approach leads to a separation in 4 quadrants using 2 thresholds on $S\hat{N}R_{post}$ and $S\hat{N}R_{prio}$.

choose $\alpha(p,k) = \pi$ because it corresponds to the smallest resulting threshold δ and then preserve SNR values corresponding to speech whatever the phase difference between speech and noise is. This choice is natural because we cannot estimate this phase difference and consequently it leads to the less speech component suppression. However any other choice can be made for $\alpha(p,k)$.

Let the *a priori* SNR threshold, η, be equal to -6dB, in this case the *a posteriori* SNR threshold, δ, is equal to nearly -6dB. This particular choice, based on experiments, is illustrated in Fig. 3. The two thresholds separate the SNR features into four quadrants (two in dark gray dots and two in light gray). The interest of this separation is the possibility to classify the features into different categories. By processing output signals using the *a posteriori* SNR values of each quadrant, informal listening tests confirm that a classification can be made. The right dark gray features lead to high level musical noise only and the ones in the two left quadrants lead to very low and inaudible components that are consequently useless. Finally the right light gray features can be classified as SNR components leading to speech components only, without musical noise. We can emphasize that a reliable classification is obtained because the behaviors of the *a posteriori* and *a priori* SNR estimators are complementary. Actually, the *a posteriori* SNR estimator is efficient for speech components but poor for musical noise and the *a priori* SNR estimator of the DD approach is efficient for musical noise but biased for speech components. As a consequence, an efficient separation of the SNR features can be done in the space defined by the 2-tuple $(S\hat{N}R_{post}, S\hat{N}R_{prio})$.

Based on this classification, we propose to re-estimate the *a posteriori* SNR using only the reliable features and to use it to compute the spectral gain. This algorithm called RFSNR is described as follows

step 1: The *a posteriori* and *a priori* SNRs are computed using relations (3) and (4), respectively.

step 2: The *a posteriori* SNR is re-estimated as follows

$$S\hat{N}R^{thr}_{post}(p,k) = \begin{cases} S\hat{N}R_{post}(p,k) & \text{if } S\hat{N}R_{post}(p,k) \geq \delta \\ & \text{and} \\ & S\hat{N}R_{prio}(p,k) \geq \eta, \\ 1 & \text{else,} \end{cases} \quad (11)$$

where thr indicates that the *a posteriori* SNR is processed using thresholds.

step 3: This re-estimated and unbiased SNR, $S\hat{N}R^{thr}_{post}(p,k)$, is directly used to compute the spectral gain, the Wiener filter [1] for example. This gain is then applied to the noisy speech to obtain the enhanced signal. We can emphasize that the *a priori* SNR is used only to select the reliable *a posteriori* SNR features, and will not be used to compute the spectral gain as in [2] since it is biased.

step 4: Another spectral gain is computed based on *a posteriori* and *a priori* SNRs of step 1 and will be used to obtain $\hat{S}(p,k)$ needed in step 1 for the next frame. Actually this is what is done in the classical DD approach.

Notice that the two right quadrants in Fig. 3 correspond to the case where a threshold is applied only to the *a posteriori* SNR values in a way close to spectral subtraction [4] and that the dark gray features are those who introduce the musical noise in the enhanced speech. In that case, a threshold of 10dB is required to suppress all the musical noise but then all the speech components corresponding to light gray dots lying between -6 and 10dB (abscissa axis) are suppressed too. Finally, using two thresholds (11) avoids this problem and allows to preserve all the features corresponding to speech components while suppressing the musical noise.

5. RFSNR BEHAVIOR ILLUSTRATION

In this example, the spectral gain chosen for the DD and RFSNR approaches is the classical Wiener gain [1]. Figure 4 shows three spectrograms. Figure 4.(a) represents the noisy speech corrupted by car noise (SNR=12dB) and Fig. 4.(b) is the enhanced speech, free of musical noise, obtained with the RFSNR technique and Fig. 4.(c) is the musical noise successfully removed.

This musical noise corresponds only to the right dark gray and the left features of Fig. 3 which confirms that the proposed features selection based on equation (11) is powerful to remove it. Notice that this very high level of musical noise is the one present in enhanced speech using only unprocessed *a posteriori* SNR (3). Furthermore, speech components are enhanced using reliable *a posteriori* SNR estimates and thus do not suffer from the bias introduced by the DD approach. Consequently the annoying reverberation effect is removed. These remarks are corroborated by informal listening tests. The remaining degradations occur because the enhancement process is based only on the amplitudes and does not take care of the phase when computing the SNRs. They also occur because the efficiency of the SNR estimators depends on the quality of the noise PSD estimation.

6. CONCLUSION

In this paper, we proposed and analyzed a new SNR estimator based on the selection of the most reliable *a posteriori* SNR features. The *a posteriori* SNR estimator is efficient for speech components but leads to high level musical noise. That is why the DD approach is preferred to compute the *a priori* SNR which efficiently reduces the level of musical noise. However, this estimator is biased for speech components leading to degradation for the enhanced speech and to an annoying reverberation effect. The complementary behaviors of these two estimators precisely allow to classify the features in the space defined by the 2-tuple $(S\hat{N}R_{post}, S\hat{N}R_{prio})$ since reliable and unreliable features are well separated. Finally, the enhanced speech is free of musical noise and does not suffer from the bias above-mentioned since only the reliable *a posteriori* SNR features are used to compute the spectral gain. Consequently, the reverberation effect characteristic of the DD approach is also removed.

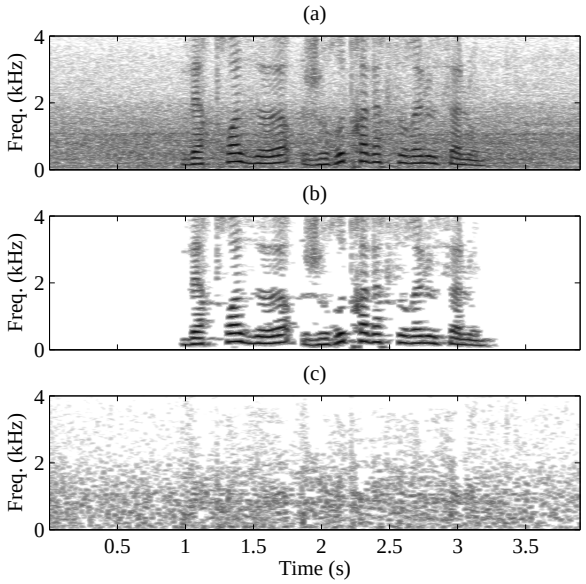

Figure 4: Speech spectrograms. (a) Noisy speech; (b) Noisy speech enhanced by RFSNR technique; (c) Musical noise successfully removed using the RFSNR technique.

7. REFERENCES

[1] P. Scalart, and J. Vieira Filho, "Speech Enhancement Based on a Priori Signal to Noise Estimation," *IEEE ICASSP'96*, Vol. 2, pp. 629–632, 7–10 May 1996.

[2] Y. Ephraim, and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on ASSP*, Vol. ASSP-32, No. 6, pp. 1109–1121, Dec. 1984.

[3] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. on SAP*, Vol. 2, No. 2, pp. 345–349, Apr. 1994.

[4] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on ASSP*, Vol. ASSP-27, No. 2, pp. 113–120, Apr. 1979.

[5] J.S. Lim, and A.V. Oppenheim, "Enhancement and Bandwith Compression of Noisy Speech," *IEEE Proc.*, Vol. 67, No. 12, pp. 1586–1604, Dec. 1979.

[6] P. Renevey, and A. Drygajlo, "Detection of Reliable Features for Speech Recognition in Noisy Conditions Using a Statistical Criterion," *Proceedings of Workshop on CRAC*, Aalborg, Denmark, pp. 71-74, 2 Sept. 2001.